LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

UCRL-JRNL-214691

Using Decision Trees for Comparing Pattern Recognition Feature Sets

D. D. Proctor

August 18, 2005

Astrophysical Journal Supplement Series

Disclaimer

Using Decision Trees for Comparing Pattern Recognition Feature Sets

Deanne D. Proctor
Lawrence Livermore National Laboratory
7000 East Ave.
Livermore, California 94550
E-mail: ddproctor@llnl.gov

Abstract

Determination of the best set of features has been acknowledged as one of the most difficult tasks in the pattern recognition process. In this report significance tests on the sort-ordered, sample-size normalized vote distribution of an ensemble of decision trees is introduced as a method of evaluating relative quality of feature sets. Alternative functional forms for feature sets are also examined. Associated standard deviations provide the means to evaluate the effect of the number of folds, the number of classifiers per fold, and the sample size on the resulting classifications. The method is applied to a problem for which a significant portion of the training set cannot be classified unambiguously.

I. Introduction

FEATURE set selection continues as a topic of research interest in pattern recognition procedure. Dasey and Micheli-Tzanakou [1] have stated that the precise choice of features is perhaps the most difficult task in pattern processing. Lam, West and Caelli [2] state, "... little research has been performed on what the best features are to use for a particular learning task." Regarding the choice of merit function, Jain, Duin, and Mao [3], in their comprehensive review of statistical pattern recognition, state that most feature selection methods use the classification error of a feature subset to evaluate its effectiveness. However, for applications in which accurately classified training sets are not available (low resolution applications, for example), recognition rates and classification errors are problematical and other approaches are necessary. In the author's initial paper [4] on the automated selection of a particular type of radio galaxy, hereafter referred to as Paper I, five, nine, fifteen and twenty-one member features sets were compared using decision trees and artificial neural networks. The lower count feature sets were subsets of the higher count sets. Adding features to the original five member feature set did not produce

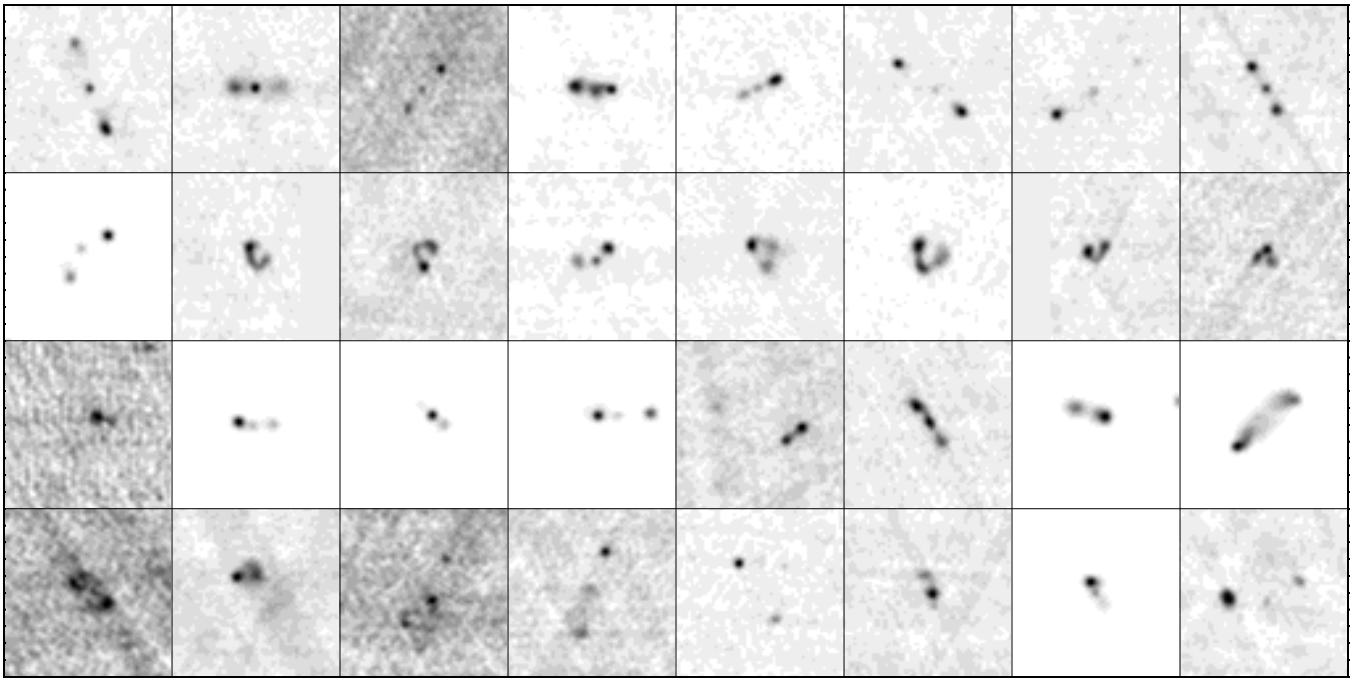

Fig. 1. Class examples: Top row, prototypical three component radio galaxies. Second row, three component bent radio galaxies. Third row, three component nonbent sources. Last row, ambiguous sources.

significantly improved solutions. As part of a process of looking for intrinsic characteristics of the target class, it is of interest to eliminate extraneous features.

In Paper I, the sort-ordered sample-size-normalized vote distribution of an ensemble of decision trees, was used to examine the ability of the decision tree classifier to generalize to previously unseen samples. This was accomplished by comparing this distribution for the training set with that of the test set. (Hereafter, the sort-ordered sample-size-normalized vote distribution will also be designated vote curve.) In this current report, vote curves are used to compare feature sets of particular interest in the application. The focus will be on comparison of feature sets using multiple runs of Oblique Classifier One (OC1), the decision tree software of Murthy, Kasif, and Salzberg [5]. More general discussions of the feature selection and evaluation process can be found in Jain and Zongker [6], Cover and Van Campenhout [7] and Narendra and Fukunaga [8].

This report is organized as follows: The background of the pattern recognition application and a summary of Paper I are presented in Section 2. Section 3 describes a series of feature set comparisons

using the sort-ordered sample-size-normalized vote distribution. Finally, Section 4 contains discussion and conclusions.

II. Background of Pattern Recognition Case Study

The pattern recognition problem under consideration is the selection of a particular type of three component radio galaxy, the so called ”bent doubles”. A proto-typical three component radio galaxy consists of two jets or lobes extending from opposite sides of a central core. Examples are shown in the first row of Fig. 1. For bent doubles the jets or lobes appear swept back as by a wind. The second row of Fig. 1 shows examples of this target class. The target class is to be separated from nonbent, S-shaped, and chance-projection three component sources. Examples of nonbent double three component sources are shown in the third row of Fig. 1. The final row of the figure shows examples of ambiguous sources, those for which visual classification is uncertain due to poor resolution or low signal-to-noise ratio. The data used in this study comes from the images and catalog [9] developed by the Faint Images of the Radio Sky at Twenty Centimeters (FIRST) Survey [10] collaboration. The catalog includes source positions, fitted parameters relating to source size and flux density and noise estimates. A random sample of 2823 sources were selected from the available population of about 15,000 three-component sources. The entire sample was visually assigned to bent double, nonbent double or the ambiguous class, the counts being N_{bent}=147, $N_{nonbent}$=1395, and N_{amb}=1281 respectively. This sample is designated the training/test set. The training set consists of only the visual bent and nonbent sources exceeding signal-to-noise ratio of 8.5, consisting of $N_{bent,tr}$=115 visual bents and $N_{nonbent,tr}$=930 visual nonbents and excludes ambiguous sources. The signal-to-noise ratio is defined as the peak flux of the component having the smallest peak flux divided by its root-mean-square error. That a significant portion of the training/test set was assigned an ambiguous classification was attributed to the relatively low resolution of the survey, 99 percent of components having fitted major axis less than 12 pixels.

One of the classifiers studied in Paper I was Oblique Classifier One (OC1). It is a system to generate a decision tree from a training set of numerical features of known classes, attempting to produce a tree that

TABLE I

LIST OF FEATURES FOR FIVE-FEATURE CLASSIFIER

1.	d_{mid}	intermediate length of pairwise distances between components
2.	d_{min}/d_{mid}	ratio of smallest distance to intermediate distance
3.	$(d_{mid}+d_{min})/d_{max}$	ratio of sum of intermediate and smallest distances to largest distance
4.	R_{SS}	ratio of silhouette sizes of assumed lobes or jets (smaller to larger)
5.	T_{SS}	total calculated silhouette size, all three components

has pure samples of training set objects. OC1's default impurity measure, the twoing rule [11], was used. (The impurity measure is the metric that is used to determine the "goodness" of a hyperplane location.)

An initial set of five basic features was used to generate classifiers and subsequently features were added. Five, nine, fifteen and twenty-one member feature sets were used. The features used were all derived from catalog entries of the three components. Table 1 gives the features used for the five-feature classifiers. Distances are projected distances on the plane of the sky. The geometry is illustrated in Fig. 2. The core is assumed to be the component opposite the longest leg of the triangle formed by the three components, the other components being possible lobes or jets. The silhouette sizes are calculated by evaluating the number of pixels with flux density greater than a threshold for a model calculated from the catalog entries of the component.

Cross validation was used, with the training/test set being divided into five folds. The training set members from four folds were used to classify the entire remaining fold, each fold thus being classified in succession from the classifiers generated by the other four folds.

The OC1 search algorithm includes some randomization to avoid local minima in the search space. Heath, Kasif, and Salzberg [12] have shown the accuracy of classification is improved by having multiple trees vote. Thus, for each feature set ten classifiers were generated for each of the five folds. For accurate

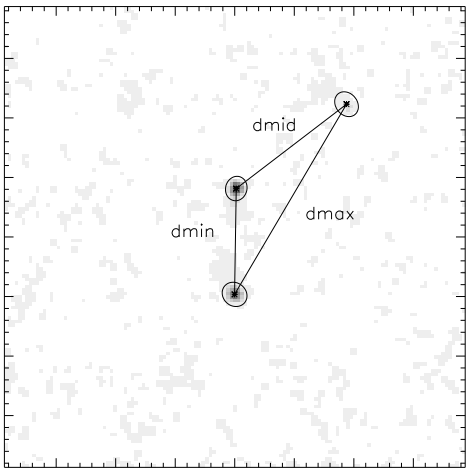

Fig. 2. Projected geometry, three component source. The core is assumed opposite side of size dmax, the other sources are presumed lobes or jets or chance projections.

classification and adequate features, it is expected the unpruned decision tree when acting on the training data would produce $N_{bent,tr}$ sources classified as target and $N_{nonbent,tr}$ sources classified as nontarget. Typically decision trees are pruned to avoid overfitting of the data. Details of OC1 pruning can be found in Reference 5. Each tree's vote on a source was apportioned using the prescription followed by White et al. [13] for pruned decision trees. Using this prescription, if a sample ends up at a leaf node with N training set objects of which B are bent, the tree's single vote on the source is split into the fraction (B+1)/(N+2) in favor and the fraction (N-B+1)/(N+2) against bent classification. The votes of the ten trees in favor of each source were then averaged. It is this normalized score, shown in subsequent comparison plots, that provides an estimate of the probability of individual three-component source being of the target class.

When the five, nine, fifteen and twenty-one feature classifiers were compared, recognition rates and false positives were within about one mean square error of each other using a somewhat arbitrary top 16% of the vote curve being classified bent. In this paper we present more extensive comparisons. The vote curves for some specific feature set comparisons are examined in the following section. Only the

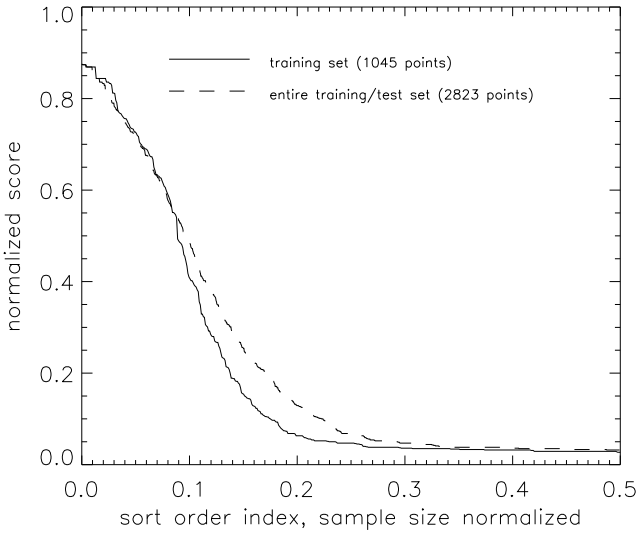

Fig. 3. Vote curve comparison of training set with entire training/test set, five feature classifier.

issue of comparing feature sets with OC1 decision tree vote distributions will be addressed.

III. SOME FEATURE SET COMPARISONS

The above discussion leads to the expectation that, ideally, the area under the training set vote curve should be $N_{bent,tr}/(N_{bent,tr} + N_{nonbent,tr})$ and thus constant for a given training set. This allows evaluation of feature sets based on compactness of the vote curve as well as the comparison of vote distributions for the visual bents. Examples follow a brief discussion of generalization.

Generalization is the ability of a classifier to classify previously unseen samples. Here and in the previous report, it was examined by comparison of the training set vote curve with the entire training/test set vote curve. This implicitly assumes the ambiguous portion of the population has same distribution in feature space as nonambiguous population. Depending upon the application this may or may not be a reasonable assumption.

Fig. 3 shows the comparison, for the five-feature classifier, of the training set vote curve and the training/test set vote curve. (Since the distributions were essentially flat after normalized index 0.5, only

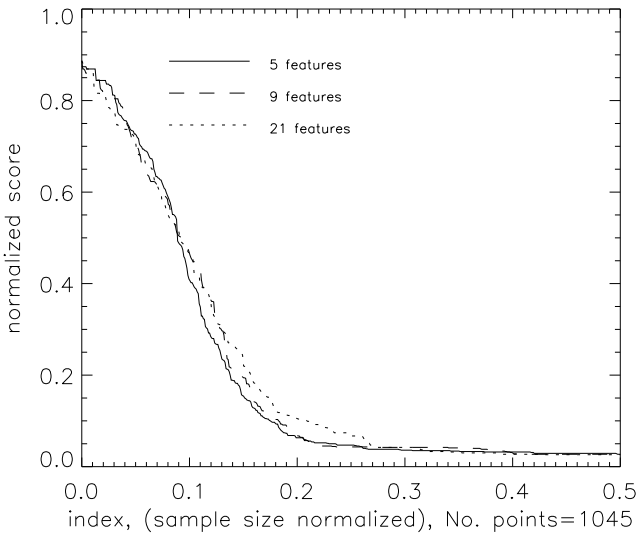

Fig. 4. Vote curve comparison of five, nine, and twenty-one feature classifiers (training set). Fifteen feature vote curve intermediate between nine and twenty-one feature vote curve.

initial half of distribution is shown.) Consistent with the 15 feature comparison shown in the Paper I, the five feature comparison suggests fairly good generalization from training to test set. In Fig. 3 the area under the training sample curve is 0.112, compared with the bent fraction ($N_{bent,tr}/(N_{bent,tr}+N_{nonbent,tr})$ = 0.110). This compares with the area under the curve for the entire training/test set of 0.125, an approximately 11% difference. Ideally the distributions would overlap, starting at 1 and dropping vertically to 0 at the true, but for this application, unknown, target fraction. It should also be noted that this ideal may not be attainable due to lack of sufficiently distinguishing features to break the degeneracy.

A. Comparison of Five, Nine, Fifteen and Twenty-one Member Feature Sets

Fig. 4 shows a comparison of the vote curves of the training set for five, nine, and twenty-one feature classifiers, whereas Fig. 5(a) shows the curves for the entire training/test set. While the distributions in Fig. 5(a) are less compact, overall the relative order of the feature sets is the same for both figures. The fifteen feature classifier distribution was intermediate between the nine and twenty-one feature classifier

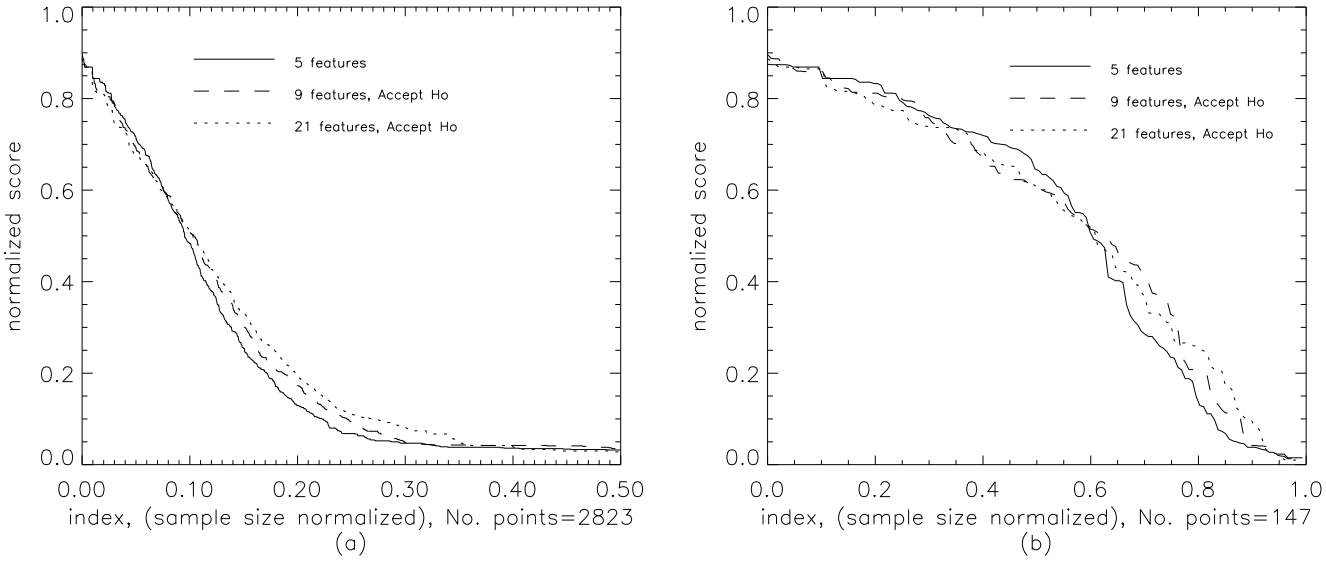

Fig. 5. Vote curve comparisons of five, nine, and twenty-one feature classifiers, (a) entire training/test set (b) visual bents.

for both figures and was omitted to improve plot clarity. In this comparison, the five feature set distribution appears the most compact, and thus the most desirable.

In this and following comparisons, the results of statistical tests at the 5% significance level are reported. For the visual bent vote curve, as Fig. 5(b), the Kolmogorov-Smirnov [16] test, comparing two cumulative distribution functions, and the Wilcoxon signed rank test [16], comparing effects of two treatments on paired data, were applied. If the statistical test results were in agreement, the mutual result is reported, if not, the results are listed in order Kolmogorov-Smirnov result and Wilcoxon signed rank test result. For the training/test set vote curves, as in Fig. 5(a), score values above 0.05 are compared using Conover's distribution functions [17] for Tsao's [18] truncated Smirnov statistics. Details and discussion of this selection for the training/test set are in Section 3.5. In all instances, the null hypothesis is H_o: no difference in distributions under consideration.

The significance tests show that, at the 5% significance level, when each of the nine, fifteen, and twenty-one feature training/test set vote curves is compared with the five feature vote curve, Fig. 5(a), the

hypothesis of equivalent distributions is accepted. Fig. 5(b) shows corresponding vote curves for visual bents only. The significance tests show that at the 5% significance level the hypothesis of equivalent distributions is accepted. These results appear consistent with noise introduced by inclusion of extraneous features causing slight degradation in the compactness of the vote distributions for the entire sample, but the classifier being able to generate substantially equivalent classifications for the visual bents.

Fig. 6 is a direct comparison of the vote of the five feature classifier with the vote of the twenty-one feature classifier for each training point. A small random offset was added to improve visualization. While there is relatively good agreement on most very low scoring sources (normalized vote less than 0.05 for both classifiers), there is considerable scatter in higher vote sources. Correlation coefficients between the five and 21 feature classifier votes for the entire bent/nonbent training set is 0.88, whereas, for visual bents alone, the correlation coefficient is 0.79.

Examination of vote variance as a function of vote value shows smallest variance at extreme vote values, the variance being larger for mid-range values.

For each of the subsequent comparisons, distributions for the training set showed the same relative order as the entire training/test set distributions. Thus for subsequent comparisons, only the results of the entire training/test set will be shown.

B. Comparison of Five-Member Feature Set with its Various Four-Member Feature Subsets

Since Fig. 4 and Fig. 5 suggest no substantial benefit from adding features to the original five member feature set it is of interest to look at feature sets with fewer members. As noted by Kittler [15] and demonstrated above, ”Redundant and irrelevant information has derogatory effect on classification process.” From a data-mining viewpoint, interest is in determination of intrinsic characteristics of the target class. Interpretation of decision tree results is difficult with even as few as three features, since the number of decision trees per feature set is the number of folds times the number of trees per fold. Though resulting classifications may be similar, interpretation of results is simpler without extraneous features. Thus, in the interest of reducing the five member feature set, decision trees were attempted dropping each of the

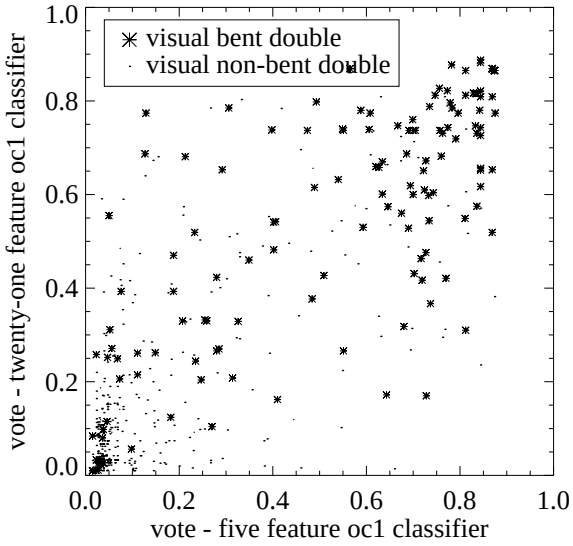

Fig. 6. Vote comparison of five and twenty-one feature classifiers. Number of points=1542. Eleven of 147 visual bent doubles had both classifier scores less than 0.05. Most points are clustered in lower left corner.

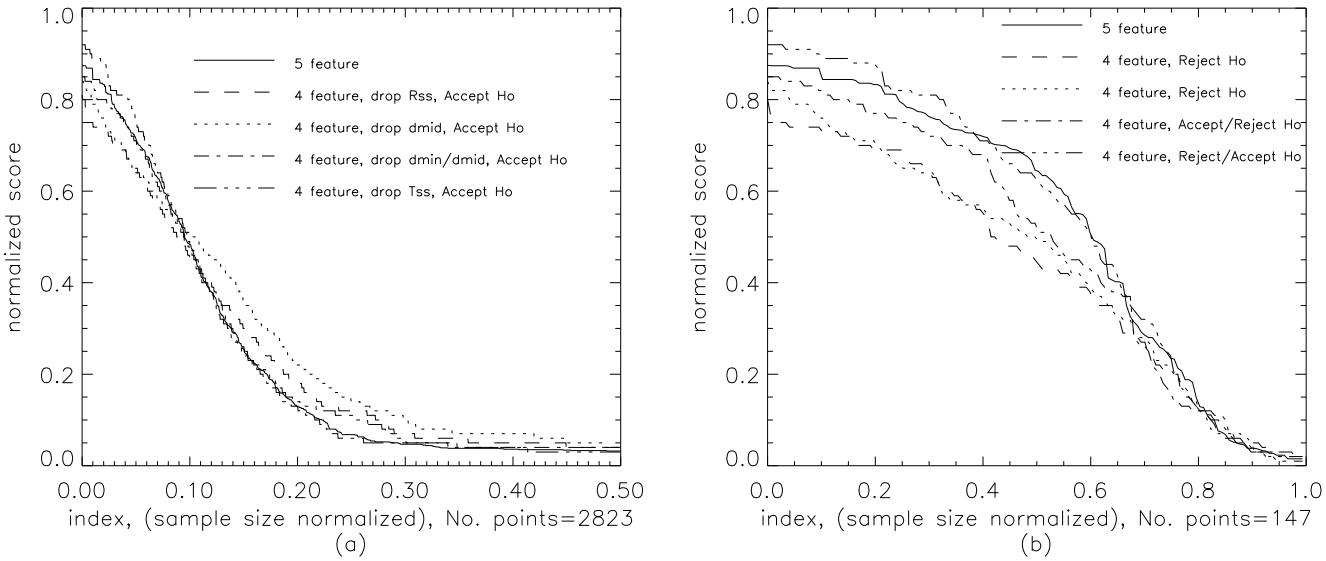

Fig. 7. Vote curve comparisons of five feature classifier with its various four feature classifier subsets, (a) entire training/test set, (b) visual bents. The excluded feature is listed in (a). OC1 was not successful in separating classes when $(d_{mid}+d_{min})/d_{max}$ was dropped.

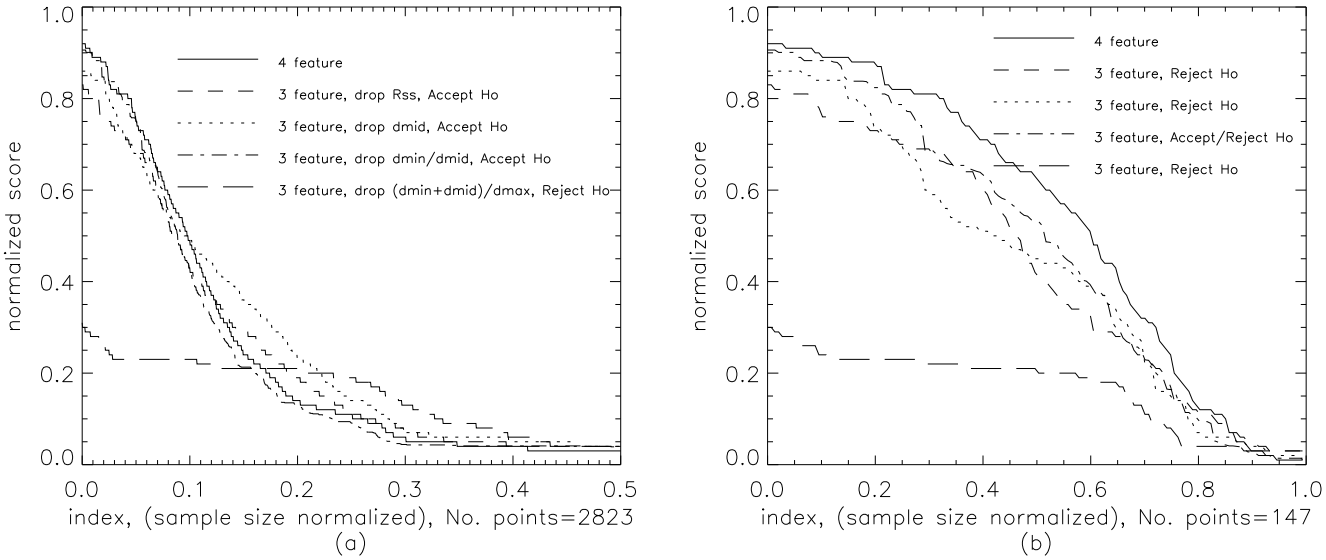

Fig. 8. Vote curve comparisons of four feature classifier with its various three feature classifier subsets, (a) entire training/test set, (b) visual bents.

five features of Table 1 in succession. Fig. 7 shows the vote distribution comparison of these feature sets. The feature being dropped is indicated in the legend in part (a) of the figure. OC1 was not successful in separating classes when the bentness ratio, (d_{mid}+d_{min})/d_{max}, was dropped from the five feature set. The significance test results show the training/test set vote curves for the successful four feature classifiers are not significantly different from the five feature classifier at the 5% level. As shown in Fig. 7(b), dropping R_{SS} and d_{mid} resulted in significantly different and degraded visual bent vote curves, indicating necessity of these members of the feature set, whereas dropping d_{min}/d_{mid} and T_{ss} showed mixed results. Since, dropping total silhouette size, T_{SS}, produced the more compact curve, whereas dropping the other features resulted in degraded distributions, T_{SS} will be excluded in remining comparisons.

C. Comparison of Four-Member Feature Set with its Various Three-Member Feature Subsets

To examine even simpler feature sets, decision trees were attempted dropping, in succession, each of the four features of previous best four feature set. A comparison of the vote distributions are shown in

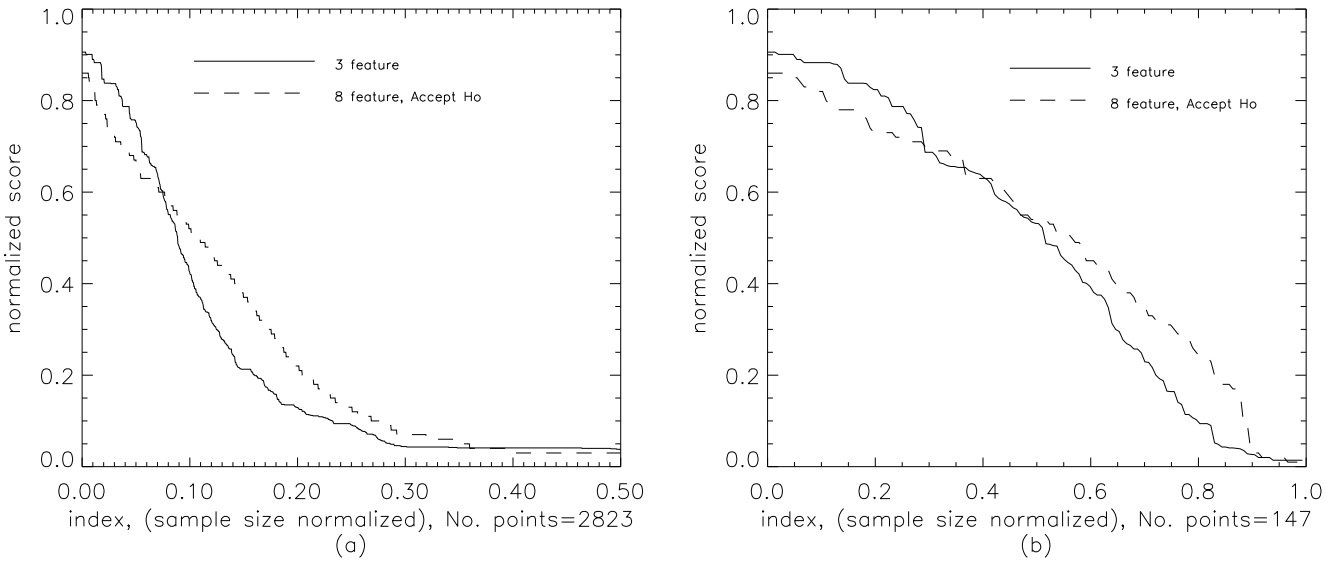

Fig. 9. Vote curve comparisons of three feature classifier with expanded-form eight feature classifier, one of three features expanded in terms of its six components, (a) entire training/test set, (b) visual bents.

Fig. 8. Again, the legend in part (a) of the figure indicates the dropped feature. Dropping the projected arm length ratio, d_{min}/d_{mid} has the least effect on the training/test set vote curve, whereas dropping the bentness ratio, (d_{mid}+d_{min})/d_{max}, has the most deleterious effect. Dropping R_{SS} and d_{mid} have intermediate effects. Significance test results are as shown. As for the four feature classifiers, features d_{mid}, (d_{mid}+d_{min})/d_{max} and R_{SS} are needed, with d_{min}/d_{mid} of perhaps more marginal necessity. In further comparisons, d_{min}/d_{mid} will be dropped as a feature.

D. Alternative Forms for Variables

At this point it is of interest to compare classifications resulting from the best three feature set (d_{mid}, (d_{mid}+d_{min})/d_{max}, R_{SS}) with an eight feature set (d_{mid}, (d_{mid}+d_{min})/d_{max}, six constituent catalog variables of R_{SS}), R_{SS} being the ratio of silhouette sizes of assumed jets or lobes, smaller to larger. This comparison examines the ability of the classifier to deal with complex relationships. The fitted model functional form

of the flux density S(x,y) at position (x,y) is given by

$$\mathrm{S}(\mathrm{x},\mathrm{y}) = \mathrm{Sp}\exp(-(\frac{x^2}{2\sigma_x^2} + \frac{y^2}{2\sigma_y^2})), \tag{1}$$

where Sp, σ_x and σ_y are derived from catalog entries for the component. The number of pixels greater than a threshold is then calculated to determine the silhouette size of the component and the appropriate ratio taken for R_{SS}.

The vote curves for this feature set comparison are shown in Fig. 9. As might be expected, the three feature training/test set distribution appears more compact, though it is not significantly different at the 5% significance level. The hypothesis tests show at the 5% level, the visual bent distributions are equivalent. This is a rather powerful example of the ability of the decision tree classifier to adapt to different functional forms of the features, assuming all relevant information is available. There is again considerable scatter in the direct vote comparison for the visual bent doubles (not shown).

A second alternative-forms comparison is for the three feature set (R_{SS}, d_{mid}, (d_{mid}+d_{min})/d_{max})) compared with the four feature set (R_{SS}, d_{min}, d_{mid}, d_{max}). These comparisons are shown in Fig. 10. Here, the visual bent vote curve is nearly identical for the two forms and the scatter is somewhat reduced in the direct vote comparison (not shown).

Though statistical tests indicate differences are not significant at the 5% level, in both alternative forms cases the more compact feature set was associated with the more compact training/test set distribution, as might be expected if it is the ratios that are of significance, not particular magnitudes of the features. None the less, the visual vote curves were equivalent at the 5% significance level. However, there is still considerable scatter in the direct vote comparison for the higher vote sources, though less so in the four feature comparison than the eight feature comparison.

E. Classifier Generation Comparison

In order to examine the sensitivity of the vote to decision tree generation, a separate five-fold, ten-classifiers-per-fold, decision tree ensemble was generated using different random number seeds for the

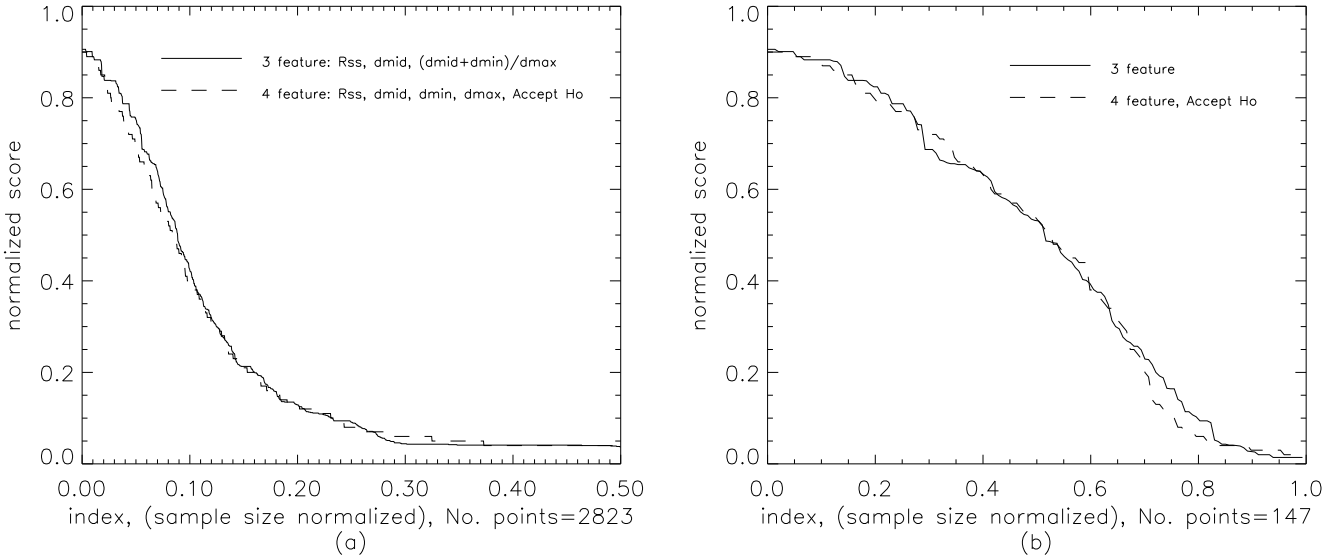

Fig. 10. Vote curve comparisons of three feature classifier with four feature expanded-form classifier, (a) entire training/test set, (b) visual bents.

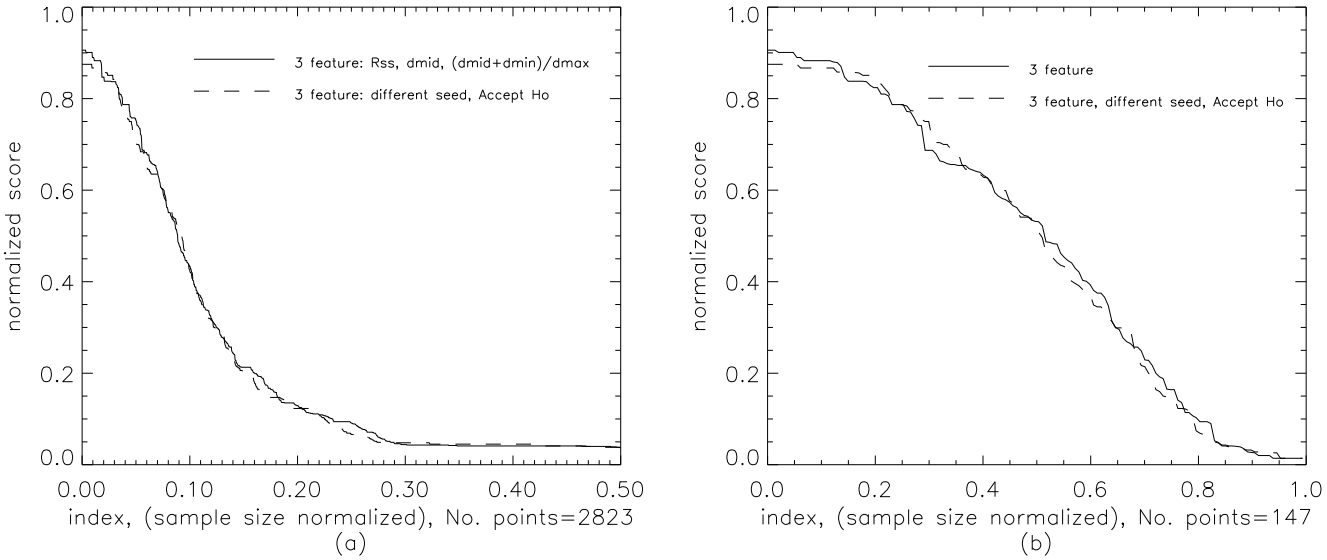

Fig. 11. Vote curve comparisons of two separate generations of three feature classifier, (a) entire training /test set, (b) visual bents.

above best three feature classifier. These results are compared in Fig. 11. Note the continuous interweaving of the distributions in Fig. 11, in contrast to previous comparisons. Initial hypothesis tests using Kolmogorov-Smirnov and Wilcoxon signed rank tests on curves in Fig. 11(a) resulted in rejection of the hypothesis of equivalent distributions, clearly not the expected result. This rejection appears to be an artifact of the relatively small number of folds and the quantization of decision tree results, there being large numbers of a few small but slightly different values for the two generations. Since details of the vote curves below say 0.05 are not of particular interest, the curves above that value were compared using Tsao's truncated Smirnov's distribution [18] as developed by Conover [17]. A random sample of 60 points from each training/test set was examined. Using this statistic, the hypothesis of equivalent visual vote curves is accepted. In Fig. 12, the direct vote comparison, the higher vote sources show better agreement than previous cases, suggesting classifier generation using five folds with ten classifiers per fold is a less significant source of error than the feature set selection. Direct vote comparison with 20 initializations per fold, five fold classifiers and 10 initializations per fold, 20 fold classifiers showed similar scatter, suggesting feature set selection or visual classification a larger source of error than classifier generation. Examination of the scatter in the classifications of a training set of half size showed similar variation to the full training set, again suggesting visual classification and inadequacy of feature set the largest source of error. Comparison of the vote curves for half-size training set classifiers with full-size training set classifiers showed non-significant differences at the 5

F. Two Member Feature Sets

Next, the various two feature subsets of the above best three feature set are compared in Fig. 13. Again dropping the bentness ratio, $(d_{mid}+d_{min})/d_{max}$, has the most significant impact. Dropping d_{mid} has an intermediate effect, and dropping R_{SS} has smaller, though significant, effect. The significance tests on the visual bent vote curves reiterate the necessity for all three features.

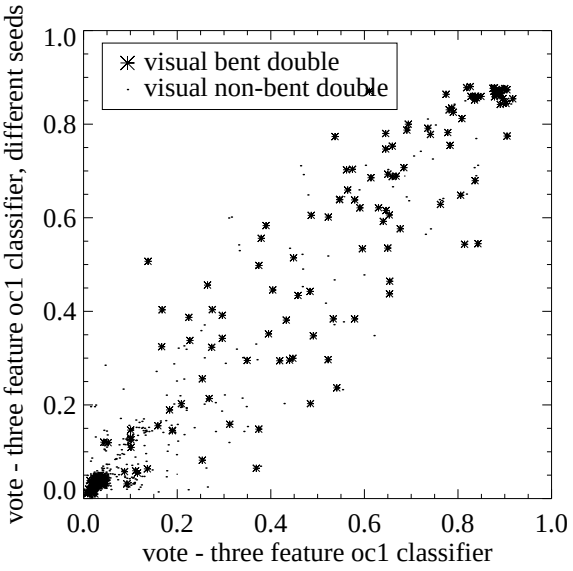

Fig. 12. Vote comparison of two separate generations of three feature classifier.

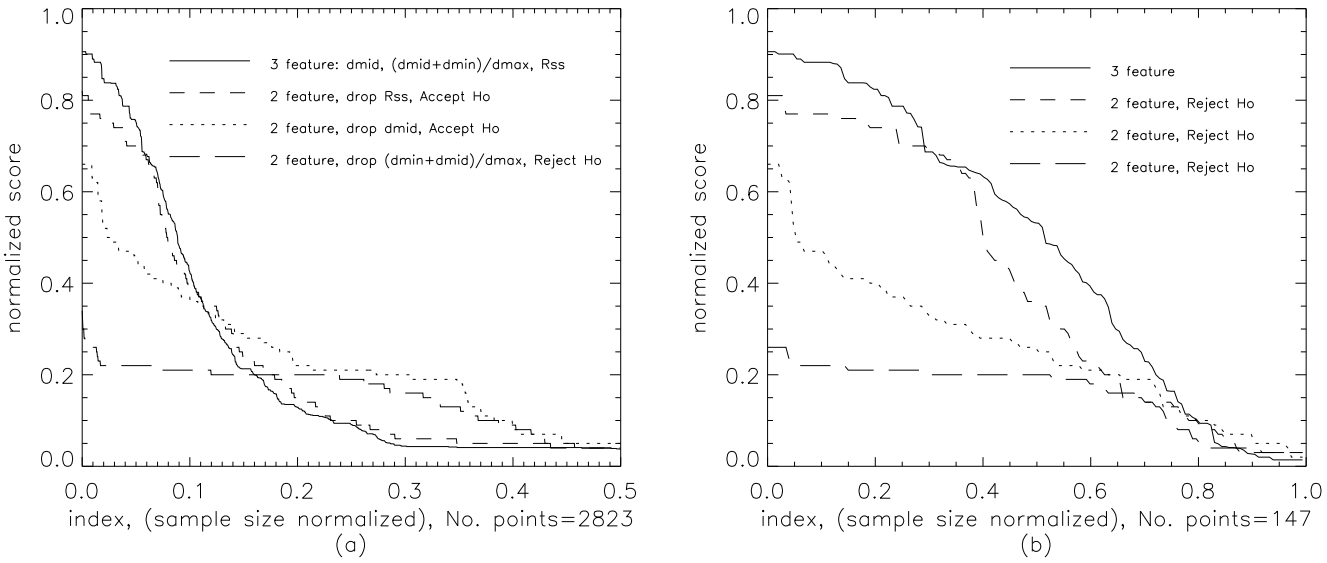

Fig. 13. Vote curve comparisons of three feature classifier with its various two feature classifier subsets, (a) entire training/test set and (b) visual bents.

G. Feature Space Plots and Result Comparison

For the above best three feature classifier, as an alternative to detailed examination of the fifty decision trees in the ensemble, two dimensional visualization can be employed to deduce the region of feature space occupied by the target class. Fig. 14 and Fig. 15 show plots of d_{mid} vs. (d_{mid}+d_{min})/d_{max} for various R_{SS} intervals. Fig. 14 shows the visual bent and nonbent classifications, while Fig. 15 shows sources with vote greater than .5 as bold. Overall results are as expected, in that the target class has higher bentness ratio and ratio of silhouette sizes closer to one. However, best boundary values would have been difficult to determine without pattern recognition algorithms. It is noted that re-examination of sources classified as bent in the two top plots of Fig. 14 suggest they may be some of the more dubious visual classifications.

Finally Fig. 16 shows 32 highest ranked sources (vote value =0.86) from the best four feature classifier. These can be compared with Fig. 17 showing 32 randomly selected lowest ranked sources (vote value =0.03) from that classifier. Results seem consistent with respective estimated probabilities.

IV. Discussion and Conclusions

Specific feature set comparisons have been demonstrated using the sort-ordered, sample-size-normalized vote distribution of an ensemble of decision trees. While recognition rates and classification errors may be adequate for feature set comparison in some applications, the the sort-ordered, sample-size-normalized vote distribution appears to provide a more comprehensive method for this application, where the determination of recognition rates and classification errors are problematical due to the uncertainty in the visual classification.

A case was observed where dropping a feature resulted in somewhat improved compactness of the vote distribution. Dropping the total silhouette size, T_{ss}, from the five feature set, demonstrated marginal improvement with deletion of a feature.

Results of the alternative-forms comparison are as expected, in that the more compact, lower-count feature set produces the more compact vote curve and demonstrate the ability of the classifier to handle

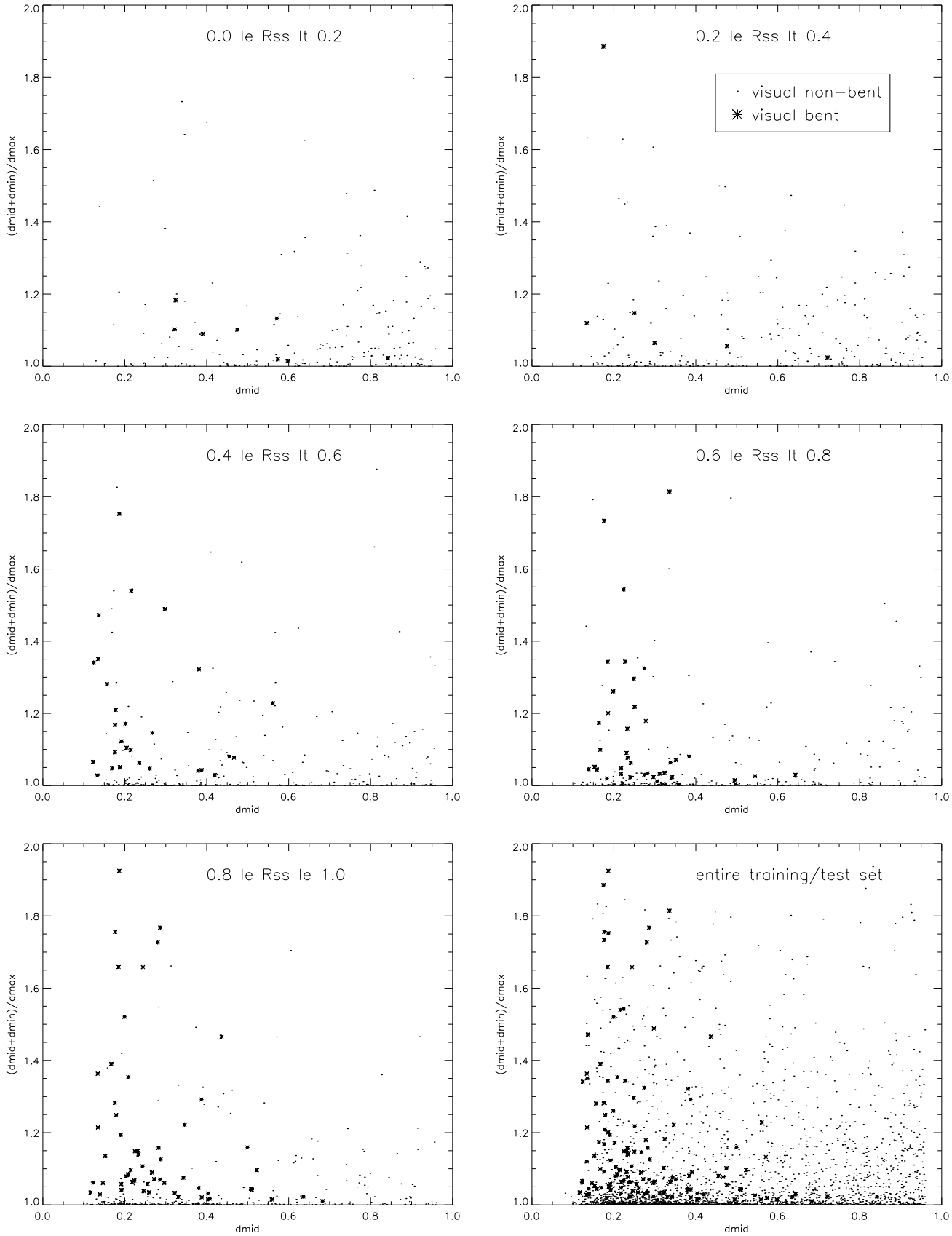

Fig. 14. Visual bent and nonbent sources as function of d_{mid} and $(d_{mid}+d_{min})/d_{max}$ for various R_{SS} ranges.

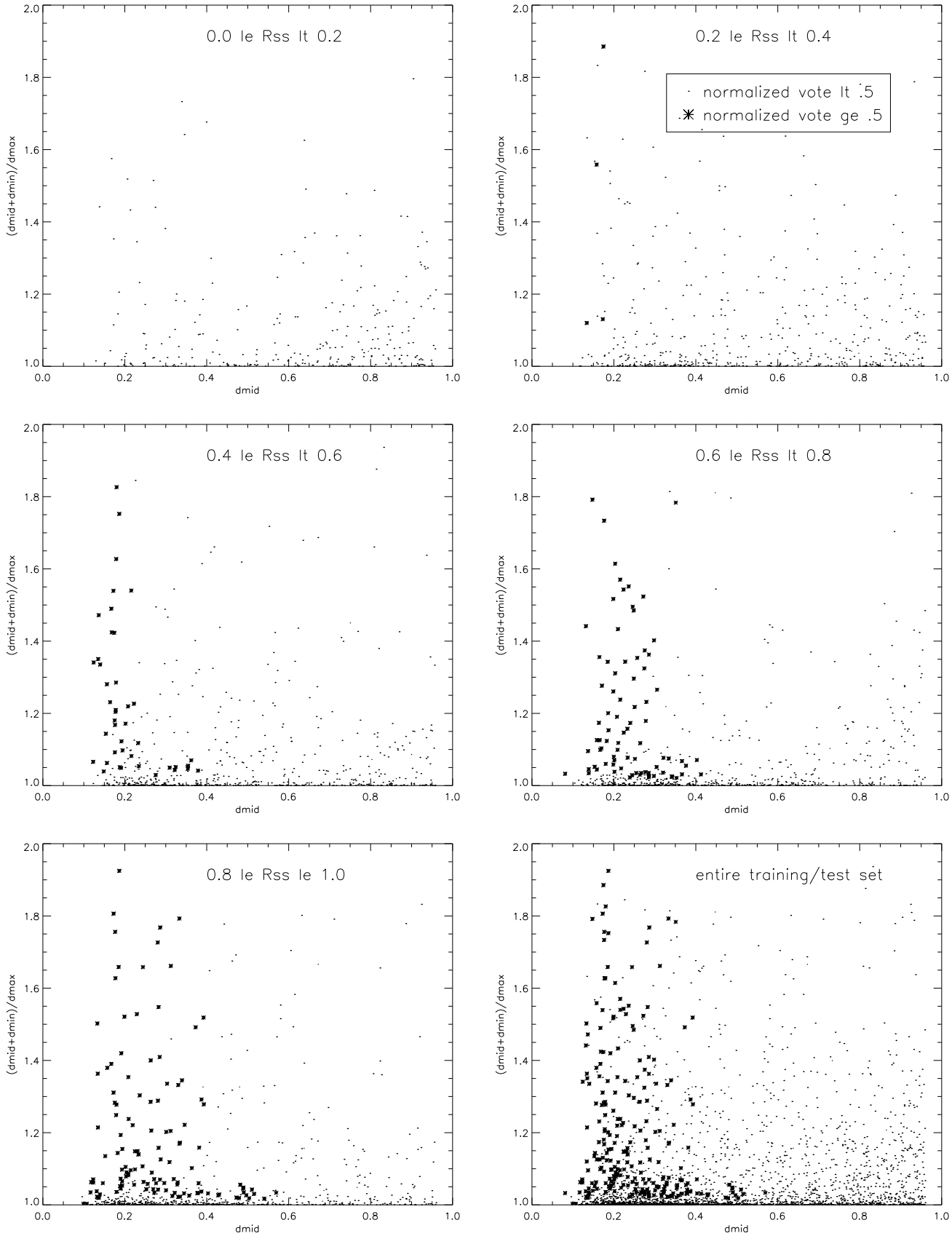

Fig. 15. Vote comparison as function of d_{mid} and $(d_{mid}+d_{min})/d_{max}$ for various R_{SS} ranges.

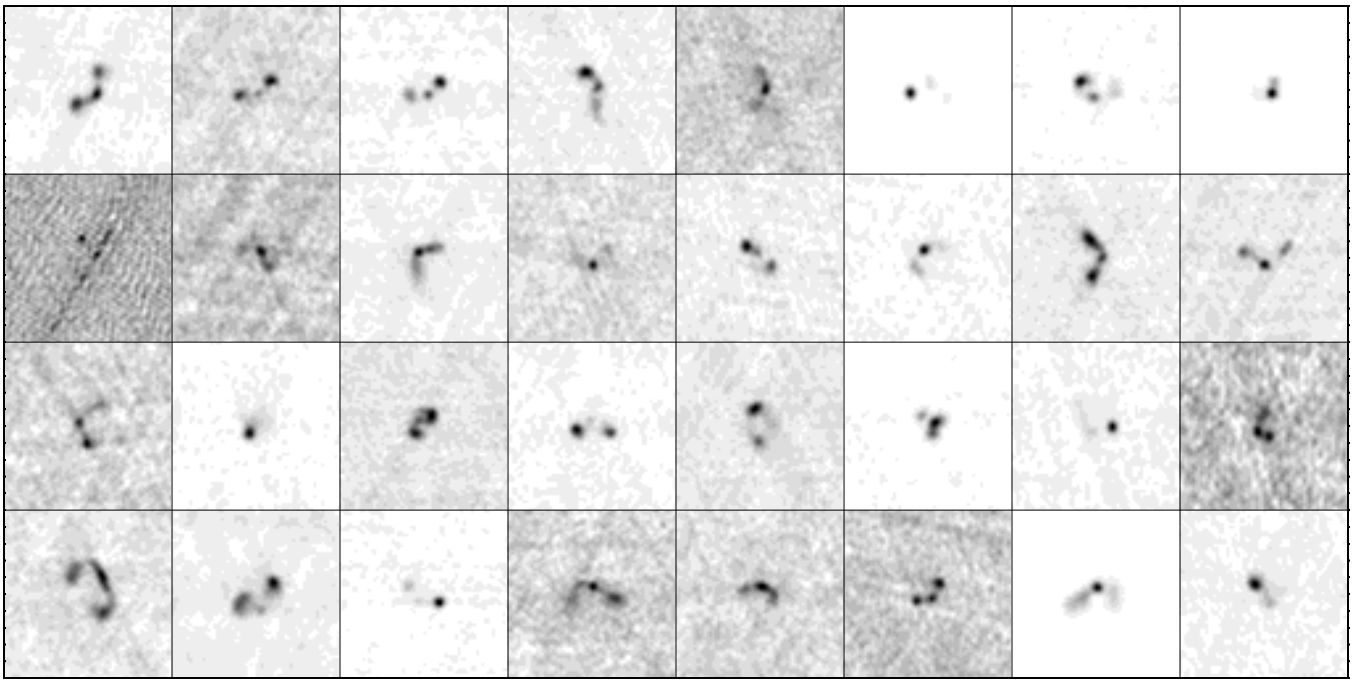

Fig. 16. The 32 highest ranked sources from best four feature classifier.

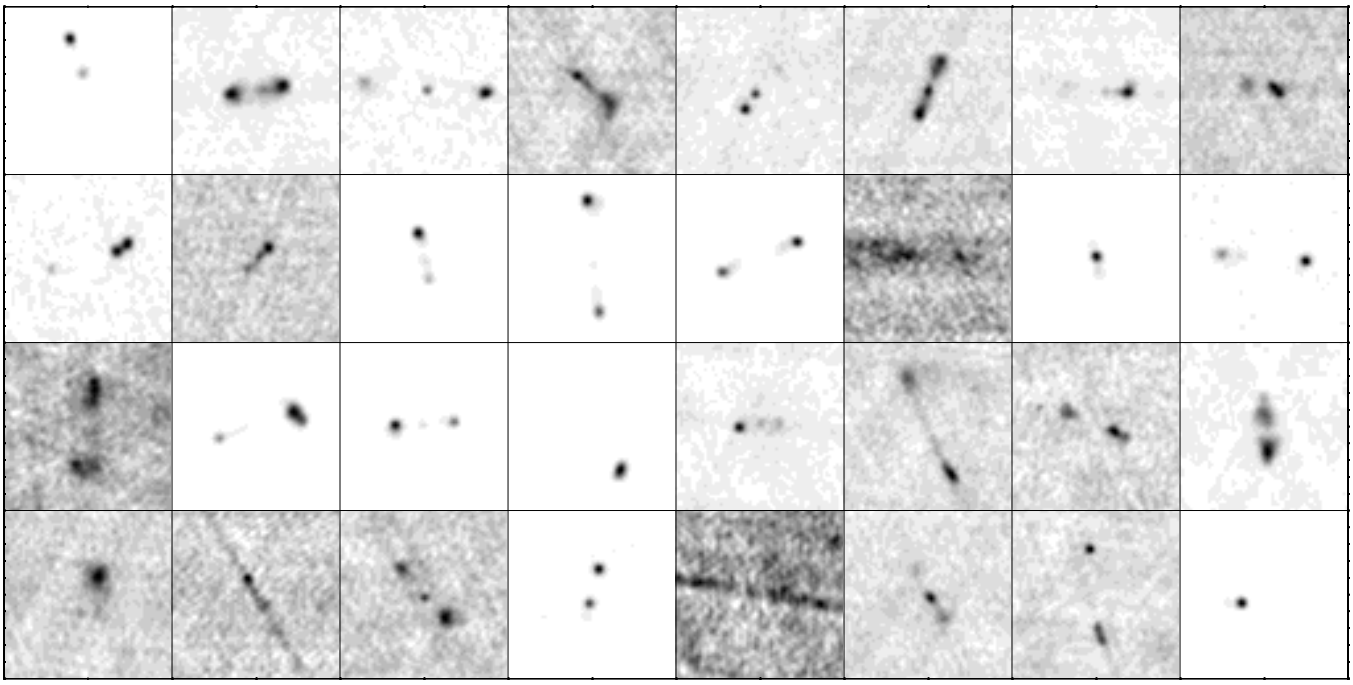

Fig. 17. Random selection of 32 of lowest ranked sources from best four feature classifier.

complex functional relationships. It is possible for some applications that expanded forms could produce better classifiers.

Of the feature sets examined, the four feature set, d_{mid}, $(d_{mid}+d_{min})/d_{max}$, R_{SS}, d_{min}/d_{mid}, provided the most desirable visual bent vote distribution, though the d_{min}/d_{mid} feature is of arguable necessity. It is noted the optimal feature subset may not have been found. It is expected that would require exhaustive search or application of branch and bound [8] techniques.

Vote curve analysis provides a method to evaluate the effect of training set size, number of folds and number of classifiers per fold on classification errors. Using multiple classifiers per fold allows error estimation on the probability of sample being of the target class. While OC1 was the particular decision tree system used in this study, the method would be applicable to otherdecision tree systems employing randomization in generation of the classifiers.

ACKNOWLEDGMENT

R. Becker provided computer resources. Richard White provided software to access the FIRST images as well as discussion of vote apportionment for pruned decision trees. The author is greatly appreciative of office space and computing facilities provided by the Institute of Geophysics and Planetary Physics (IGPP), John Bradley and Kem Cook.

The term 'vote curve' was coined by an anonymous referee.

This work was performed under the auspices of the U.S. Department of Energy, National Nuclear Security Administration by the University of California, Lawrence Livermore National Laboratory under contract No. W-7405-Eng-48.

REFERENCES

[1] T. J. Dasey, E. Micheli-Tzanakou in *Supervised and Unsupervised Pattern Recognition*, E. Micheli-Tzanakou, Ed., pp. 135-162, CRC Press, New York, NY (2000).

[2] C. P. Lam, G. A. W. West, T. M. Caelli, "Validation of machine learning techniques, decision trees and finite training set," *J. Electron. Imaging* 7, 94-103, (1998).

[3] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: a review", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1),4-37 (2000).

[4] D. D. Proctor, "Low-resolution pattern recognition - sorting triples in the FIRST database," *J. Electron. Imaging* 12, 398-409 (2002).

[5] S. K. Murthy, S. Kasif and S. Salzberg, "A system for induction of oblique decision trees," *J. Artif. Intell. Research*, 2, 1 (1994).

[6] A. Jain, D. Zongker, "Feature selection: evaluation, application, and small sample performance", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 19, 2, 153-158 (1997).

[7] T. M. Cover and J. M. Van Campenhout, "On the possible orderings in the measurement selection problem," *IEEE Trans. Systems, Man, and Cybernetics* 7, no. 9, 657-661, (1977).

[8] P. M. Narendra and K. Fukunaga, "A branch and bound algorithm for feature subset selection," *IEEE Trans. on Computers*, C-26, 9, 917-922 (1977).

[9] R. L. White, R. H. Becker, and D. J. Helfand, M. D. Gregg, "A catalog of 1.4 GHz radio sources from the FIRST survey," *Astrophys. J.* 475, 479-493 (1997).

[10] R. H. Becker and R. L. White, D. J. Helfand, "The FIRST survey: faint images of the radio sky at twenty-cm.," *Astrophys. J.* 450, 559-577 (1995).

[11] L. Breiman, J. Friedman, R. Olshen, C. Stone, *Classification and Regression Trees*, Wadsworth International Group (1984).

[12] D. Heath, S. Kasif, S. Salzberg, *Cognitive Technology: In Search of a Humane Interface*," ed. B. Gorayska & J. Mey Elsevier, Amsterdam, 305 (1996).

[13] R. L. White, R. H. Becker, M. D. Gregg, S. A. Laurent-Muehleisen, M. S. Brotherton, C. D. Impey, C. E. Petry, C. B. Foltz, F. H. Chaffee, G. T. Richards, W. R. Oegerle, D. J. Helfand, R. G. McMahon, and J. E. Cabenela, "The FIRST bright quasar survey. II. 60 nights and 1200 spectra later," *Astrophys. J. Supp. Series* 126, 133-207 (2000).

[14] W. Press, S. Teukoisky, W. Vettering, B. Flannery, *Numerical Recipes in C, Second Edition*, Cambridge University Press, Cambridge (1992).

[15] J. Kittler, "Feature selection and extraction", *Handbook of Pattern Recognition and Image Processing*, Academic Press, New York, (1986), Chap. 3, 59.

[16] B. Ostle, *Statistics in Research, 2ed*, Iowa State University Press, Ames, Iowa (1964).

[17] W. J. Conover, "The distribution functions of Tsao's truncated Smirnov statistics", p. 1208-1215.

[18] C. K. Tsao, "An extension of Massey's distribution of the maximum deviation between two sample cumulative step functions," Ann. Math. Statist. 25, 587-592 (1954).

List of Tables

List of Figures